APLICACIÓN DE *SKOS* PARA LA INTEROPERABILIDAD DE VOCABULARIOS CONTROLADOS EN EL ENTORNO DE *LINKED OPEN DATA*

Juan-Antonio Pastor-Sánchez, Francisco-Javier Martínez-Méndez y José-Vicente Rodríguez-Muñoz

Juan-Antonio Pastor-Sánchez es doctor en documentación y profesor de la *Facultad de Comunicación y Documentación* de la *Universidad de Murcia*, en el área de construcción de servicios de información digital. Lleva a cabo su investigación en el ámbito de las tecnologías de la web semántica y la gestión de contenidos digitales.

Universidad de Murcia
Facultad de Comunicación y Documentación
Campus Universitario de Espinardo, 30100 Murcia
pastor@um.es

Francisco-Javier Martínez-Méndez es doctor en documentación por la *Universidad de Murcia* y profesor de tecnologías de la información en la *Facultad de Comunicación y Documentación*. Principalmente desarrolla su investigación en el área de la recuperación de la información en la Web.

javima@um.es

José-Vicente Rodríguez-Muñoz es doctor en informática y catedrático del *Área de Biblioteconomía y Documentación* de la *Universidad de Murcia*. Su actividad científica se ha centrado en los campos de la gestión de información, recuperación de información y evaluación de sistemas de búsqueda web.

jovi@um.es

Resumen

Se pretende determinar el alcance de la aplicación de *SKOS* en el ámbito de la web semántica. Para ello se han analizado los vocabularios controlados que aplican *SKOS* y que se encuentran registrados en *The data hub*, un catálogo de conjuntos de datos RDF. Además de establecer una tipología de vocabularios, se analizan los aspectos relacionados con el acceso abierto a los datos mediante su descarga directa, su consulta a través de un *Sparql endpoint* y la existencia de licencias adecuadas. Se examina el nivel de interoperabilidad existente a través de la definición de relaciones de mapeado entre vocabularios. Se concluye indicando que tesauros y clasificaciones están más integrados en el entorno de los datos abiertos, mientras que los encabezamientos de materia poseen un mayor grado de interoperabilidad.

Palabras clave

SKOS, Vocabularios controlados, Clasificaciones, Tesaurus, Web semántica, *Linked open data*, Interoperabilidad semántica.

Title: *SKOS* application for interoperability of controlled vocabularies in the field of *linked open data*

Artículo recibido el 29-02-12
Aceptación definitiva: 23-05-12

Abstract

This paper aims to determine the scope of application of SKOS in the Semantic Web. Controlled vocabularies registered in the *The Data Hub* catalog of RDF datasets that use SKOS were studied. In addition to establishing a typology of vocabularies, authors analyzed the issues related to open access to data through its direct download, its query through a *Sparql endpoint* and the existence of appropriate licenses. The existing level of interoperability through the definition of mapping relationships between vocabularies are examined. The study concludes that thesauri and classifications are more integrated into the open data environment while subject headings lists have greater interoperability.

Keywords

SKOS, Controlled vocabularies, Classifications, Thesaurus, Semantic web, Linked open data, Semantic interoperability, *Sparql endpoint*.

Pastor-Sánchez, Juan-Antonio; **Martínez-Méndez, Francisco-Javier**; **Rodríguez-Muñoz, José-Vicente**. "Aplicación de *SKOS* para la interoperabilidad de vocabularios controlados en el entorno de *linked open data*". *El profesional de la información*, 2012, mayo-junio, v. 21, n. 3, pp. 245-253.

http://dx.doi.org/10.3145/epi.2012.may.04

Introducción

SKOS (*Simple knowledge organization system*) es una de las ontologías que mayor éxito y aplicación ha alcanzado en el entorno de la web semántica. Ofrece un modelo para representar la estructura básica y el contenido de esquemas de conceptos tales como listas encabezamientos de materia, taxonomías, esquemas de clasificación, tesauros y cualquier otro tipo de vocabulario controlado[1]. El desarrollo de *SKOS* comenzó en el seno del grupo de trabajo *SWAD-Europe* hacia el año 2002 y se difundió públicamente mediante un borrador en noviembre de 2005. En aquel momento la propuesta se dio a conocer como *SKOS Core*, etiqueta que aún perdura en muchos trabajos actuales. En agosto de 2009 *SKOS* alcanzó el estatus de recomendación del *W3C* (2009). En comparación con otras soluciones aportadas, tales como esquemas *XML ad hoc*, ontologías quizá más complejas o *topic maps*, *SKOS* ofrece una alternativa cuya aplicación es sencilla y rápida (**Pastor-Sánchez**; **Martínez-Méndez**; **Rodríguez-Muñoz**, 2009).

> En el contexto de *linked open data* el establecimiento de vínculos entre conjuntos de datos abiertos resulta esencial y los lenguajes documentales deben desempeñar un papel relevante

Actualmente son numerosos los sistemas de organización del conocimiento de todo tipo publicados mediante *SKOS* disponibles para su uso y consulta. Se han realizado breves estudios al respecto (**Isaac** *et al.*, 2011) que muestran una visión muy general de la implantación de *SKOS*. Sin embargo, creemos necesario determinar de forma más precisa su auténtico potencial, valorando su impacto, evolución y aplicación en los conjuntos de datos asociados a lenguajes documentales de todo tipo. De esta forma sabremos si esta iniciativa tiene visos de asentarse definitivamente en el ecosistema de la web semántica o si por el contrario es algo pasajero.

Algunos trabajos anteriores (**Möller** *et al.*, 2010) abordan enfoques más generales con conjuntos de datos RDF[2] sin llegar a abordar aspectos relacionados con la interoperabilidad. Únicamente analizan la explotación real de conjuntos de datos, volumen de transacciones o índices de análisis de consultas *Sparql*[3] estableciendo métricas en este sentido. Algo más específico es el estudio de **Francesconi** *et al.* (2008) que analiza las definición de vínculos de mapeado entre tesauros aplicando técnicas de recuperación de información.

El presente estudio global de la presencia de *SKOS* en la publicación de vocabularios controlados tiene como objetivo identificar su tipología, volumen y relaciones existentes con otros vocabularios o conjuntos de datos. De este modo se podrá ponderar debidamente el alcance de esta tecnología, especialmente en el ámbito de *linked open data*, contexto donde el establecimiento de vínculos entre conjuntos de datos abiertos resulta esencial y donde los lenguajes documentales deben desempeñar un papel relevante.

SKOS

SKOS se define formalmente como una ontología *OWL-full*[4] que permite representar cualquier tipo de sistema de organización del conocimiento mediante RDF. Su ámbito de aplicación se extiende a la práctica totalidad de vocabularios controlados: clasificaciones, tesauros, encabezamientos de materia, taxonomías, tesauros, glosarios, etc.

En *SKOS* los elementos de un vocabulario se representan mediante conceptos entre los que se establecen relaciones semánticas jerárquicas (simples o transitivas) y asociativas. A los conceptos se les asocian etiquetas en distintos idiomas:

- Preferentes: equivalentes a los descriptores en un tesauro. Un mismo concepto sólo puede tener una etiqueta preferente en cada idioma.
- Alternativas: similares a los no-descriptores. Permite enriquecer semánticamente un vocabulario definiendo varios puntos de acceso a un concepto.
- Ocultas: no son visibles directamente a los usuarios y se utilizan para su procesamiento por aplicaciones informáticas.

Mediante la extensión *SKOS-XL* es posible definir relaciones entre etiquetas, por ejemplo, cuando una etiqueta es un acrónimo o un préstamo lingüístico de otra.

Pueden definirse esquemas de conceptos y colecciones. Los esquemas agrupan conceptos normalmente asociados a un campo semántico o área de conocimiento determinada. *SKOS* ofrece dos propiedades que permiten relacionar un concepto con uno o varios esquemas y especificar si se sitúa como cabecera de una estructura jerárquica (*top concept*). Las colecciones permiten crear grupos de conceptos que complementan las estructuras de relaciones semánticas jerárquicas. Un mismo concepto puede formar parte de varias colecciones.

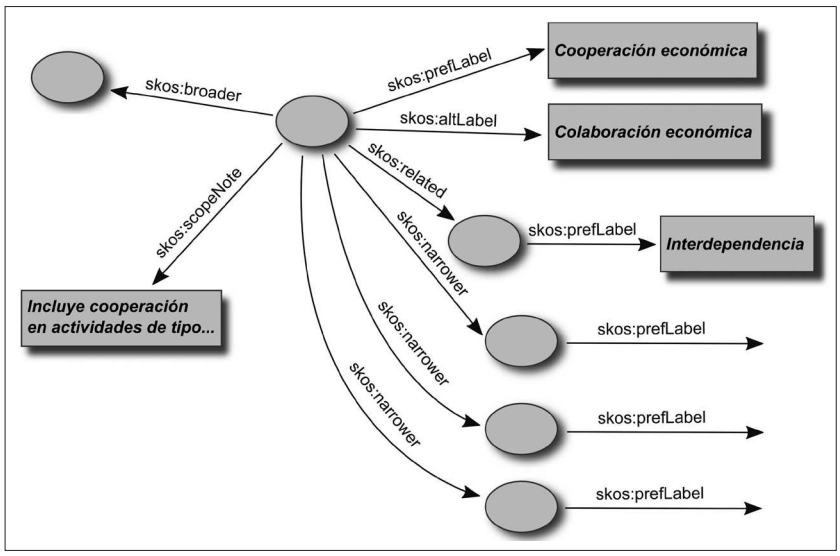

Figura 1. Representación de elementos de un vocabulario con *SKOS*
http://www.w3.org/People/Ivan/CorePresentations/SW_Advanced/Slides.html#(73)

SKOS ofrece una serie de relaciones semánticas para establecer vínculos de mapeado entre conceptos de diferentes esquemas. Esto permite indicar si un concepto de un esquema se considera idéntico a otro o cuándo tienen un significado cercano, genérico, específico o relacionado. La nueva norma de tesauros *ISO 29562* propone una función similar para definir relaciones entre diferentes lenguajes documentales con el objeto de poder utilizarse conjuntamente en operaciones de recuperación de información.

> La relaciones de mapeado de *SKOS* son clave para la participación de los sistemas de organización del conocimiento en SRI basados en la web semántica

Por ejemplo, un repositorio digital utiliza un vocabulario controlado X para indizar documentos y un banco de imágenes hace lo propio con un vocabulario Y. Entre ambos vocabularios podrían definirse correspondencias entre los diferentes conceptos que lo componen. A través de una consulta en el banco de imágenes se recuperarían elementos del repositorio digital y viceversa. Estas relaciones son claves para la participación de los sistemas de organización de conocimiento en el escenario *linked open data*, puesto que el grado de interoperabilidad de un vocabulario controlado se determina por el establecimiento de relaciones de mapeado con otros vocabularios o conjuntos de datos.

The data hub como fuente de referencia del estudio

La fuente principal para la toma de datos de este estudio es el catálogo en línea *The data hub*[5] (*TDH*). Se trata de un catálogo colaborativo con información descriptiva sobre todo tipo de conjuntos de datos disponibles en internet, mantenido por la *Open Knowledge Foundation* como parte de una iniciativa destinada a dar soporte a actividades para la difusión del conocimiento de carácter abierto.

La información que ofrece *TDH* se refiere a la identificación y localización de conjuntos de datos, descripción del tipo de contenido, ejemplos, información sobre la posibilidad de descarga total o parcial de dichos conjuntos de datos, existencia de *Sparql endpoints*[6], número de tripletas RDF, vínculos con otros conjuntos de datos, espacios de nombres utilizados, etc. El servicio también permite asignar a cada conjunto de datos diferentes etiquetas que describen el contenido y formatos utilizados.

Se seleccionaron los conjuntos de datos que representan vocabularios controlados de todo tipo mediante *SKOS*. Debido al carácter abierto y cooperativo de este catálogo, las descripciones de los conjuntos de datos son heterogéneas y a menudo incompletas e incorrectas. Por ello, en algunas ocasiones se recuperaron resultados que, si bien en la descripción o etiquetas incluían el término "SKOS", en realidad no se referían a ningún vocabulario controlado o realizaban un uso deficiente o incorrecto de *SKOS*. Se verificó si existían conjuntos de datos adicionales referidos a tesauros, taxono-

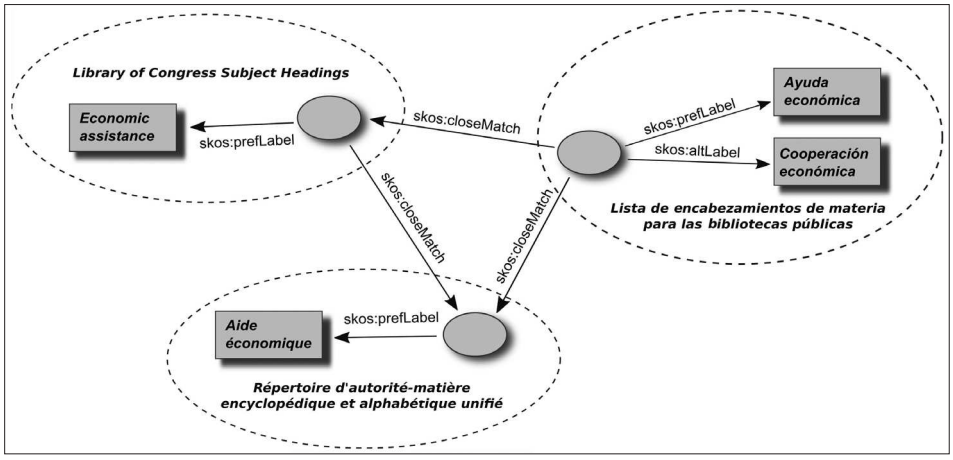

Figura 2. Interoperabilidad de vocabularios entre las *LEM*, *LCSH* y *Rameau*

mías, clasificaciones y glosarios de los que, pese a no aparecer descritos bajo la etiqueta "SKOS", pudiera verificarse que hacen uso de dicha especificación. Esta verificación tuvo resultado negativo[7].

> Mientras que los tesauros son el tipo de vocabulario controlado más frecuente sobre el que se aplica *SKOS*, el mayor volumen de datos (más de un 73%) corresponde a los ficheros de control de autoridades

Se consultaron dos fuentes suplementarias: la página sobre *datasets* de *SKOS* del *W3C*[8] y el informe del *Grupo Incubadora* (**Isaac** *et al.*, 2011) sobre conjuntos de datos y vocabularios controlados en el ámbito de *library linked data*. Gracias a la primera fuente se localizaron 3 vocabularios adicionales: los del servicio *NASA Taxonomy*, el *UK Archival thesaurus* (*UKAT*) y la representación en *SKOS* de las distintas categorías de *Wikipedia* del proyecto *Dbpedia*. Otros vocabularios mencionados en dicha página no han podido ser analizados debido a la imposibilidad del acceso conjunto a los datos correspondientes o por hacer referencias a proyectos o trabajos inconclusos[9].

De cada vocabulario se recopiló la siguiente información:

- denominación;
- volumen;
- relaciones con otros conjuntos de datos (incluidos los correspondientes a otros vocabularios);
- disponibilidad de servicios de consulta mediante *Sparql endpoint*;
- descarga de ficheros con datos RDF en cualquier serialización;
- existencia de licencias compatibles con el uso libre y abierto de los datos.

Esta información no pudo ser obtenida para todos los vocabularios a través de la búsqueda en *TDH* debido a las deficiencias del catálogo mencionadas anteriormente. Fue necesario recopilarla por otros medios, siempre y cuando fuera viable el acceso, descarga y análisis de los datos correspondientes. Se analizaron los ficheros fuente mediante el validador de documentos RDF del *W3C*[10], obteniendo la información del número de tripletas y de vínculos con datos externos. La obtención de la información sobre la disponibilidad de los datos para su descarga o consulta mediante *Sparql endpoint* ha requerido la navegación por los sitios web donde están publicados los vocabularios. Algunos de ellos, como los encabezamientos de materia suecos o los vocabularios de la *Biblioteca Nacional de Hungría*, han quedado fuera de este estudio porque *TDH* no ofrece información estadística sobre el volumen de datos y relaciones y tampoco fue posible acceder y descargar los conjuntos de datos correspondientes.

Resultados obtenidos

Los resultados se han dispuesto en varios grupos de características para un análisis estructurado. Se han estudiado los tipos de vocabularios identificados para una comprensión más adecuada del contexto de aplicación de *SKOS* en el ámbito de *linked open data*. También se ha analizado el nivel de apertura de los vocabularios para su consulta y descarga, teniendo en cuenta la presencia explícita de una licencia compatible con la reutilización abierta y libre de los datos. Finalmente se ha examinado el grado de interoperabilidad de los vocabularios a través de las relaciones de mapeado establecidas entre sí[11].

Tipos de vocabularios y volumen de datos

Se han examinado conjuntos de datos correspondientes a 55 vocabularios controlados agrupados en 6 tipos: tesauros, listas de encabezamientos de materia, ficheros de control de autoridades, clasificaciones, léxicos y vocabularios, taxonomías y ontologías. Como refleja la tabla 1, más de la mitad de los conjuntos de datos analizados se corresponden con tesauros (19) y clasificaciones (16). Desde el punto de vista del volumen de datos, de los más de 347 millones de tripletas que forman los conjuntos de datos analizados, un 73,6% (tabla 1) se corresponden con ficheros de control de autoridades y un 10,6% con listas de encabezamientos de materia. En consecuencia, la distribución porcentual del número de vocabularios en función de su tipo (% frecuencia) no se corresponde con la distribución porcentual de tripletas de cada tipo (% volumen de datos).

Los vocabularios de mayor volumen son el *Fichero de autoridades virtual internacional* (*VIAF*) y la *Aplicación facetada de terminología temática* (*FAST*) basada en la *Lista de encabezamientos de materia de la Biblioteca del Congreso*. Resulta interesante destacar que del volumen total de tripletas de tesauros, más del 70% se corresponde con el *Tesauro de medio ambiente multilingüe general* (*Gemet*).

El "triángulo *open data*"

Hay que analizar si estos vocabularios cumplen 3 criterios que indiquen si se encuentran integrados en la iniciativa *linked open data*, es decir determinar si en cada conjunto de datos:

1. Es posible la descarga conjunta de todo el vocabulario en un formato compatible con RDF tal como *N3*, *Turtle* o *RDF/XML*. El acceso íntegro a los conjuntos de datos en alguno de estos formatos resulta esencial para su reutilización.

Tipo de vocabulario	Vocabularios analizados	Frecuencia (%)	Número de tripletas	Volumen de datos (%)
Tesauros	19	34,6	28.309.759	8,2
Clasificaciones	16	29,1	6.891.148	2,0
Encabezamientos de materia	7	12,7	36.724.495	10,6
Autoridades	4	7,3	255.528.450	73,6
Léxicos y glosarios	4	7,3	388.729	0,1
Ontologías	3	5,5	5.003.228	1,4
Taxonomías	2	3,6	14.542.505	4,2
Totales	**55**	**100**	**347.388.314**	**100**

Tabla 1. Distribuciones absolutas y porcentuales de las frecuencias y volumen de datos por tipo de vocabularios analizados

2. Existe la posibilidad de utilizar un *Sparql endpoint* para la consulta selectiva de los datos RDF correspondientes al vocabulario. De este modo es posible una reutilización más selectiva, integrando servicios web y reutilizando vocabularios.

3. Se ha establecido una licencia compatible con el uso abierto y libre de los datos.

Estos criterios se complementan entre sí: una licencia *open data* sirve de poco si no se suministran los datos o si no es posible su reutilización inmediata y selectiva a través de consultas *Sparql*. Los servicios de descarga y consulta tampoco resultan muy útiles si es imposible utilizar dichos datos debido a que carecen de una licencia adecuada (**Alexander *et al.***, 2009).

Resulta útil representar estos datos agrupándolos según el tipo de vocabulario. La tabla 2 muestra los datos absolutos en términos totales y porcentuales (columnas "abs n" y "abs %" respectivamente) del grado de cumplimiento de cada uno de los tres criterios indicados anteriormente por parte de cada tipo vocabulario. Las licencias *open data* más utilizadas en los diferentes conjuntos de datos de los vocabularios analizados son *Open knowledge definition*, *Open database licence* y diversas modalidades de *Creative commons*.

> "Clasificaciones y tesauros destacan por su alto grado de integración en el ecosistema de datos abiertos, con mecanismos y licencias adecuados para ello"

Es necesario tener en cuenta el peso de cada vocabulario en función del volumen de datos que aporta para cada tipo de vocabulario. Un vocabulario con 1.000 tripletas no tiene el mismo peso que otro que aporte 10 millones. Por este motivo se ha calculado el grado de cumplimiento de cada uno de los criterios para cada tipo de vocabulario realizando una ponderación en función del número de tripletas que aporta.

Dicho cálculo se define de la siguiente forma: sea una familia *T* de conjuntos de tripletas de cada vocabulario de un determinado tipo $\{X_1 \ldots X_n\}$ donde el número de conjuntos que forman parte de dicha familia viene definido por una función de cardinalidad[12] *card(T).* Sea X_i un vocabulario tal que $X_i \in T$ cuyo número de tripletas viene dado por su cardinalidad $card(X_i)$ y $p(X_i, C)$ una función binaria que indica si dicho vocabulario cumple (1) o no (0) un determinado criterio *C* de participación en *open data*. El nivel de participación en términos porcentuales, *OD(T,C)*, de un tipo de vocabulario *T*, en un determinado criterio *C open data*, se calcula como:

$$OD(T,C) = \sum_{k=1}^{card(T)} \left(\frac{card(X_k) \cdot p(X_k, C)}{S(T)} \cdot 100 \right)$$

donde

$$S(T) = \sum_{j=1}^{card(T)} card(X_j)$$

Tras realizar los cálculos se han obtenido los datos porcentuales (columnas $OD(T,C_x)$ de la tabla 2) mediante la ponderación en función del volumen de los vocabularios de cada tipo. El mismo cálculo se ha realizado, entendiendo que todos los vocabularios forman parte de un único conjunto, lo que nos permite obtener una visión global.

Se ha escogido una gráfica radial (figura 3) para representar el grado porcentual de cumplimiento de cada uno de los criterios para tesauros, clasificaciones, listas de encabezamientos de materia y autoridades, los cuatro tipos de vocabularios que mayor volumen de datos aportan.

Cada uno de ellos conforma lo que hemos convenido denominar el "triángulo *open data*". Destaca el alto grado de adecuación tanto de tesauros como de clasificaciones con los tres criterios, mientras que son las listas de encabezamientos de materia y los ficheros de control de autoridades los que más lejos están de cumplirlos, en especial la disponibilidad de *Sparql endpoint*. Estos resultados negativos son muy pronunciados en estos tipos, puesto que los dos vocabularios más numerosos del estudio, *VIAF* y *FAST*, carecen tanto de opciones de descarga como de *Sparql endpoint*, y en el caso de *VIAF* no se ha definido de forma explícita una licencia *open data*. El mismo cálculo se ha realizado de forma global (entendiendo las tripletas de todos los vocabula-

Tipo de vocabulario	Vocabularios analizados	Criterio 1: Descarga de datos			Criterio 2: *Sparql endpoint*			Criterio 3: Licencia *open data*		
		abs n	abs %	$OD(T,C_1)$ %	abs n	abs %	$OD(T,C_2)$ %	abs n	abs %	$OD(T,C_3)$ %
Autoridades	4	3	75	22	1	25	6	2	50	16
Clasificaciones	16	15	94	94	14	88	100	14	88	100
Encabezamientos de materia	7	5	71	18	3	43	1	5	71	95
Léxicos y glosarios	4	3	75	98	1	25	12	2	50	12
Ontologías	3	3	100	100	2	67	10	2	67	10
Taxonomías	2	2	100	100	1	50	98	2	100	100
Tesauros	19	17	89	97	11	58	92	15	79	98
Totales	**55**	**48**	**87**	**33**	**33**	**60**	**18**	**42**	**76**	**36**

Tabla 2. Resultados absolutos y ponderados de aplicación de criterios *open data* agrupados por tipos de vocabularios

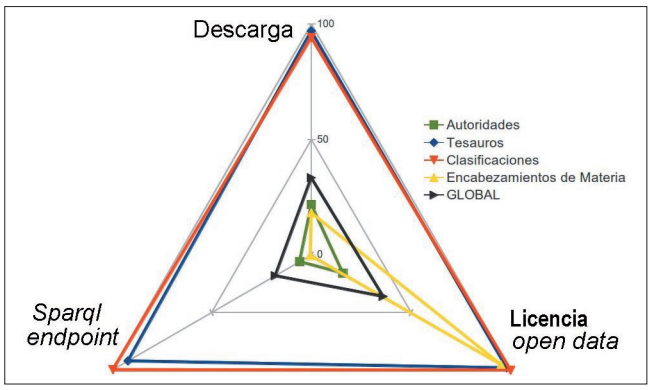

Figura 3. Triángulos *open data* de los cuatro tipos principales de vocabularios controlados

rios como parte de un único conjunto) pudiéndose observar cómo el nivel de integración en el ecosistema *open data* es muy bajo.

Interoperabilidad de vocabularios

Es otro de los aspectos que constituye un indicador de la calidad de un vocabulario (**Mader**, 2010). A mayor interconexión, mayor integración y cohesión entre conjuntos de datos. En consecuencia, se incrementan sus posibilidades de uso en sistemas de recuperación de información (SRI) basados en técnicas de la web semántica (**Angjeli** *et al.*, 2009). Esta aplicación y reutilización inmediata de los vocabularios disponibles en *SKOS* contrasta enormemente con sus equivalentes ediciones en papel, que en ocasiones han tenido una aplicación limitada y obviamente poseen una interoperabilidad limitada o nula. En *SKOS* la interoperabilidad se centra en la declaración explícita de relaciones de mapeado entre elementos de diferentes vocabularios. Por lo tanto es un camino de ida y/o vuelta: hay que considerar las relaciones mapeadas que parten de un vocabulario como las que llegan a él. Por lo tanto, si en la sección anterior se abordó en qué medida los vocabularios son *"open"*, en esta se tratará hasta qué punto podemos hablar de *"linked"*.

> Las listas de encabezamiento de materias son las que mayor visibilidad y nivel de interoperabilidad aportan en el conjunto de sistemas de organización del conocimiento representados mediante *SKOS*

Tras la recopilación de datos se han hallado 9,22 millones de relaciones: 7,75 millones (más de un 84%) corresponden a relaciones de mapeado entre vocabularios, y 1,47 millones son vínculos con otros conjuntos de datos. En esta investigación se han tratado únicamente las relaciones entre vocabularios para estudiar el nivel de interoperabilidad. Se han definido dos indicadores a los que hemos denominado *visibilidad* y *luminosidad,* utilizando la misma terminología de los estudios de la estructura de enlaces web (**Madria** *et al.*, 1999). La visibilidad se asocia a las relaciones de mapeado que un vocabulario recibe por parte de otros. Por su parte la luminosidad se refiere a las relaciones que dentro de un vocabulario se definen hacia otros vocabularios externos. El número de relaciones de mapeado se ha ponderado en función del volumen (número de tripletas) de conjunto de datos del vocabulario.

Sobre el conjunto de tripletas X_i que representa un vocabulario, y cuyo volumen se representa mediante su cardinalidad $card(X_i)$, se definen dos subconjuntos que representan las relaciones de mapeado: $R_e(X_i)$ que contiene las relaciones que el conjunto X_i recibe de otros vocabularios, y $R_s(X_i)$ con el número de relaciones de mapeado que el mismo conjunto establece con otros vocabularios externos. El número de relaciones de cada uno de los subconjuntos viene dado por las correspondientes cardinalidades $card(R_e(X_i))$ y $card(R_s(X_i))$. En consecuencia, se definen visibilidad y luminosidad de X_i, $VIS(X_i)$ y $LUM(X_i)$ respectivamente como medidas porcentuales:

$$VIS(X_i) = \frac{card(R_e(X_i))}{card(X_i)} \cdot 100$$

y

$$LUM(X_i) = \frac{card(R_s(X_i))}{card(X_i)} \cdot 100$$

También se ha definido una medida de agregación que permite estimar globalmente tanto la visibilidad como la luminosidad, aunando ambos indicadores. Esta medida a la que se ha denominado *Interoperabilidad combinada*, $IC(X_i)$, se define como:

$$IC(X_i) = \frac{card(R_e(X_i)) + card(R_s(X_i))}{card(X_i)} \cdot 100$$

equivalente a

$$IC(X_i) = VIS(X_i) + LUM(X_i)$$

Del estudio de interoperabilidad se han eliminado los vocabularios "aislados": aquellos que ni reciben, ni emiten relaciones de mapeado y para los que $IC(X_i)$ es igual a cero. La tabla 3 muestra de forma detallada y agrupada (por tipo de vocabulario) los resultados de los cálculos de luminosidad, visibilidad e interoperabilidad combinada para 31 vocabularios de los 55 que han formado parte del estudio.

Los resultados indican que las listas de encabezamientos de materia poseen un alto grado de interoperabilidad en su conjunto, debido principalmente a la alta visibilidad que tiene la *LCSH*, seguidas a gran distancia por el fichero de control de autoridades *GND* de la *Biblioteca Nacional de Alemania*. Los tesauros cuentan con una escasa interoperabilidad global, puesto que los valores de luminosidad y visibilidad alcanzados por el tesauro con mayor volumen de datos (*Gemet*) son muy bajos. La figura 4 muestra los mismos resultados de un modo más comprimido, ordenados según la medida de interoperabilidad combinada de cada vocabulario individual y tipo de vocabulario.

Vocabulario	Tipo	Abreviatura	LUM(X)	VIS(X)	IC(X)
TaxonConcept knowledge base	Taxonomía	*Taxon*	0,0001	0,0000	0,0001
Upper mapping and binding exchange layer	Ontología	*Umbel*	0,0000	0,0019	0,0019
National diet library of Japan authorities	Autoridades	*NDLNA*	0,0481	0,0000	0,0481
General multilingual enviromental thesaurus	Tesauro	*Gemet*	0,0231	0,0284	0,0515
Dewey decimal classification	Clasificación	*DDC*	0,0000	0,1017	0,1017
Common procurement vocabulary 2003	Clasificación	*CPV2003*	0,0000	0,1831	0,1831
EU's Multilingual thesaurus	Tesauro	*Eurovoc*	0,0000	0,2358	0,2358
Linked clean energy data thesaurus	Tesauro	*Reegle*	0,3840	0,0000	0,3840
Global de los tesauros		G-T	0,2720	0,1538	0,4258
Thesaurus of the FAO	Tesauro	Agrovoc	0,4151	0,1189	0,5340
Umwell Thesaurus	Tesauro	*Umthes*	1,1610	1,1610	2,3220
The Virtual international authority file	Autoridades	*VIAF*	2,0000	0,8948	2,8948
Polythematic structured subject heading system	Enc. materia	*Psshs*	3,0000	0,0000	3,0000
National Agricultural Library thesaurus	Tesauro	*NALT*	0,0000	3,6685	3,6685
Global de ficheros de control de autoridades		G-A	2,2964	2,4258	4,7222
Environmental applications reference thesaurus	Tesauro	*EARTh*	5,0000	0,0000	5,0000
Combined nomenclature 2012	Clasificación	*CN2012*	5,0035	0,0000	5,0035
Faceted application of subject terminology	Enc. materia	*FAST*	5,3333	0,0000	5,3333
Global de clasificaciones		G-C	2,4164	3,0215	5,4380
TheSoz thesaurus for the social sciences	Tesauro	Gesis	3,1584	2,8153	5,9737
Common procurement vocabulary 2008	Clasificación	*CPV2008*	0,1245	6,4135	6,5379
Thesaurus W for local archives	Tesauro	*TWLA*	9,5273	0,0000	9,5273
Répertoire d'autorité matière encyclopédique et alphabétique unifié	Enc. materia	*Rameau*	4,9817	5,7309	10,7126
International standard industrial classification V4	Clasificación	*Isicv4*	12,0299	0,0000	12,0299
Central product classification 2008	Clasificación	*CPC2008*	12,8706	0,0000	12,8706
Listas de encabezamientos de materia de las bibliotecas públicas	Enc. materia	*LEM*	13,7789	0,0000	13,7789
Gemeinsame normdatei	Autoridades	*GND*	4,6109	10,0458	14,6567
Statistical classification of products by activity 2008	Clasificación	*CPA2008*	17,3080	0,0000	17,3080
North American industry classification system 2012	Clasificación	*Naics2012*	18,3528	0,0000	18,3528
Thesaurus for graphic materials	Tesauro	T4GM	18,7379	0,0000	18,7379
North American industry classification system 2007	Clasificación	*Naics2007*	18,8278	0,0000	18,8278
Standardthesaurus wirtschaft	Tesauro	*STW*	14,9138	5,2267	20,1405
MARC codes list	Glosario	*MarcCL*	25,3176	0,0000	25,3176
Global de listas de encabezamientos de materia		G-E	4,8894	31,6983	36,5877
Library of Congress subject headings	Enc. materia	*LCSH*	1,3316	41,8306	43,1621

Tabla 3. Cálculos de visibilidad, luminosidad e interoperabilidad combinada de los vocabularios individuales y de los tipos globales

Se observa que en la mayoría de los casos la luminosidad es determinante en el cálculo de la interoperabilidad de los vocabularios: muchos de ellos con valores de cero en la visibilidad alcanzan valores de *IC(x)* que superan el 10%. En este sentido cabe señalar el alto valor de luminosidad e interoperabilidad combinada de *MarcCL* únicamente con un volumen de 8.816 tripletas. Una minoría de vocabularios se sitúan en el caso contrario: su alta visibilidad les confiere un alto grado de interoperabilidad. Tal es el caso de *GND* y *LCSH*, en especial este último cuya visibilidad es de un 41,8%, lo que permite afirmar que actualmente es el vocabulario de referencia en el entorno de *linked open data*.

Conclusiones

El catálogo *TDH* precisa una exhaustiva revisión de los datos que recoge sobre los diferentes vocabularios controlados que utilizan *SKOS*. Esta tarea podría realizarse por los propios editores de los vocabularios, conscientes de la necesidad de proporcionar una información descriptiva más homogénea y rigurosa. Los diferentes conjuntos de datos de vocabularios analizados no incorporan ficheros *VoID*[13], lo que hubiera permitido analizar detalladamente su contenido. Los catálogos colaborativos deberían registrar y

> *The data hub* resulta insuficiente para un control de los conjuntos de datos RDF: es esencial difundir el uso de ficheros *VoID* que permiten describir más detalladamente su contenido

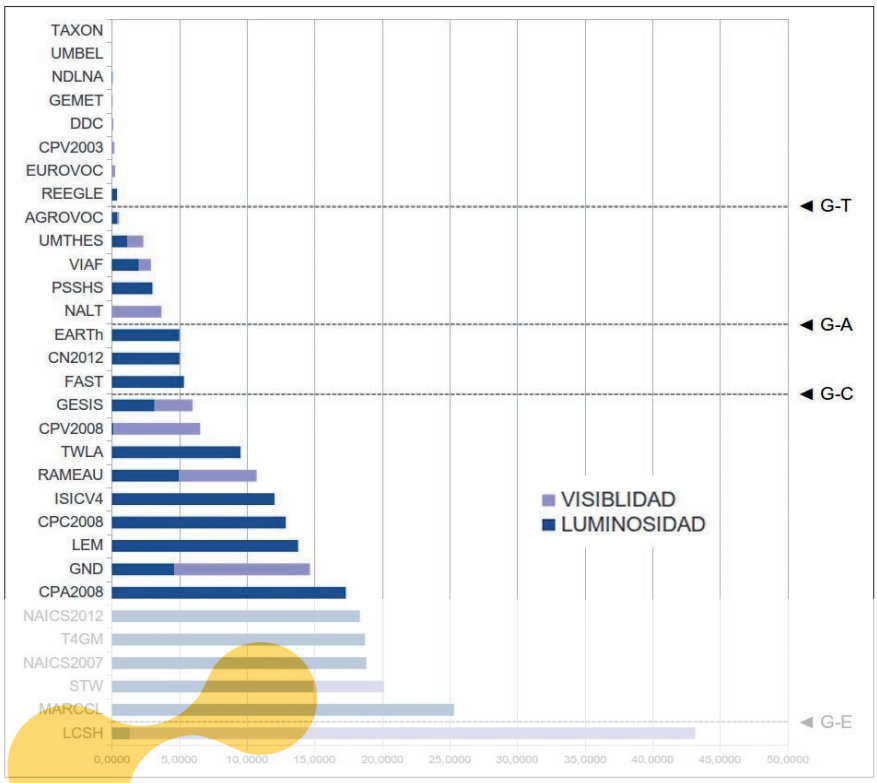

Figura 4. Interoperabilidad combinada de los vocabularios como acumulado de las medidas de visibilidad y luminosidad, y posición de la interoperabilidad global de los principales tipos de vocabularios

procesar ficheros *VoID*, sin que sea preciso registrar otros datos manualmente, puesto que la actualización de datos se realizaría inmediatamente tras un análisis automático de dichos ficheros.

La aplicación de *SKOS* en el ámbito de *linked open data* resulta muy heterogénea en función del criterio analizado y el tipo de vocabulario observado. El grado de interoperabilidad es relativamente alto en el campo de las listas de encabezamientos de materia, lo que hace pensar en un mayor nivel de penetración de *SKOS* en el sector de las bibliotecas y en los procesos de gestión y catalogación bibliográfica, que tienen como referente a los *Encabezamientos de materia* de la *Biblioteca del Congreso*. Destaca también el bajo grado de interconexión de los tesauros. El enriquecimiento de las propiedades de mapeado entre conceptos de tesauros como *Agrovoc*, *Eurovoc* o *Gemet* abriría nuevas posibilidades de reutilización de estos vocabularios y su aplicación podría conllevar un aumento de la eficiencia en SRI (no hay que olvidar que la nueva norma sobre tesauros *ISO-25964* apunta en este sentido).

Desde el punto de vista del acceso abierto a los datos, son los tesauros y las clasificaciones los que han obtenido mejores resultados, alcanzando un nivel prácticamente perfecto en todos los criterios establecidos en el estudio. Los encabezamientos de materia y ficheros de control de autoridades son muy deficientes en este aspecto, especialmente en lo referente a la existencia de *Sparql endpoints*, algo que resulta paradójico, especialmente en los encabezamientos de materia, que pese a su alto grado de interoperabilidad suelen carecer de mecanismos para su reutilización e integración. Esto puede limitar su uso, reduciendo su aplicación a las instituciones que editan dichos vocabularios.

Finalmente cabe indicar la existencia de numerosos conjuntos de datos no incorporados en este estudio, que hacen un uso incorrecto de *SKOS*. No es infrecuente encontrar conjuntos de datos que usan una única propiedad como *skos:prefLabel*, sin incorporar otro tipo de propiedades para el etiquetado de conceptos o el establecimiento de relaciones semánticas o de mapeado. Esto indica un exceso de *"skosificación"* de conjuntos de datos RDF, lo que resulta inapropiado y desaconsejable, más si cabe cuando existen otros esquemas RDF (incluso propiedades del propio RDF) que realizan esta función. La aplicación de *SKOS* debería reservarse a la publicación de sistemas de organización del conocimiento, poniendo especial énfasis en la definición de relaciones de mapeado entre diferentes vocabularios, algo que sin duda les aportará un gran valor añadido.

Notas

1. Puede encontrarse una amplia información y referencia sobre *SKOS* en: *http://www.w3.org/2004/02/skos/intro*

2. RDF: modelo marco para la descripción de recursos. Se basa en tripletas del tipo sujeto-objeto-predicado que conforman grafos complejos.

3. *Sparql* es un lenguaje de consulta para interrogar y recuperar datos RDF.

4. OWL: Lenguaje para la definición de ontologías, utilizado junto con RDF para la descripción de aspectos lógicos de las relaciones entre recursos.

5. *http://thedatahub.org*

6. Servicios web que permiten el uso de *Sparql* para realizar consultas sobre conjuntos de datos RDF y recuperarlos.

7. La búsqueda en el catálogo se realizó manejando directamente el volcado de datos de *The data hub* disponible en formato *JSON* en: *http://thedatahub.org/dump*

8. *http://www.w3.org/2001/sw/wiki/SKOS/Datasets*

9. Existe una versión de la *CDU* en *SKOS* disponible en: *http://www.udcc.org/udcsummary/exports.htm* Sin embargo se trata de una versión muy incompleta por lo que no se ha incluido en este trabajo.

10. *http://www.w3.org/RDF/Validator*

11. Los datos obtenidos para la realización de este trabajo están disponibles para su descarga en: *http://skos.um.es/files/skos-lod-2012.ods*

12. La cardinalidad de un conjunto es el número de elementos que contiene.

13. *VoID*: *Vocabulary of intelinked datasets*. Esquema RDF que permite expresar diferentes tipos de metadatos sobre conjuntos de datos RDF.
http://vocab.deri.ie/void

Bibliografía

Alexander, Keith; **Cyganiak, Richard**; **Hausenblas, Michael**; **Zhao, Jun**. "Describing linked datasets: on the design and usage of *VoiD*, the *vocabulary of interlinked datasets*. En: *Linked data on the web workshop* (*LDOW 09*), 2009.
http://events.linkeddata.org/ldow2009/papers/ldow2009_paper20.pdf

Angjeli, Anila; **Isaac, Antoine**; **Cloarec, Thierry**; **Martin, Frédéric**; **Van-der-Meij, Lourens**; **Matthezing, Henk**; **Schlobach, Stefan**. "Semantic web and vocabulary interoperability: an experiment with illumination collections". En: *IFLA International cataloguing and bibliographic control*, April-June 2009, pp. 25-29.

Francesconi, Enrico; **Faro, Sebastiano**; **Marinai, Elisabetta**; **Peruginelli, Ginevra**. "A methodological framework for thesaurus semantic interoperability". En: *Fifth European semantic web conf.*, 2008, pp. 76-87.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.140.4905&rep=rep1&type=pdf

Isaac, Antoine; Waites, William; Young, Jeff; Zeng, Marcia. *Library Linked Data Incubator Group: Datasets, value vocabularies, and metadata element sets. W3C Incubator Group Report*, 25 Oct. 2011.
http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025
http://skos.um.es/Incubator/lld/XGR-lld-vocabdataset

ISO. *ISO 25964-1:2011. Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*. ISO, 2011.

ISO. ISO/DIS 25964-1:2011. Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies. ISO, 2011.

Mader, Christian. *"Quality assurance in collaboratively created Web vocabularies". En: European semantic web conf., 2010.*
http://eprints.cs.univie.ac.at/3355/4/eswc_proposal_short.pdf

Madria, Sanjay; **Bhowmick, Sourav S.**; **Ng, Wee-Keong**; **Lim, Ee-Peng**. "Research issues in Web data mining". En: *Procs of the 1st Intl conf on data warehousing and knowledge discovery*, 1999, pp. 303-312.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.4020&rep=rep1&type=pdf

Miles, Alistair; **Bechhofer, Sean**. *SKOS Simple knowledge organization system reference*. W3C Recommendation, 18 August 2009.
http://www.w3.org/TR/skos-reference

Möller, Knud; **Hausenblas, Michael**; **Cyganiak, Richard**; **Grimnes, Gunnar-Aastrand**. "Learning from linked open data usage: patterns & metrics". En: *Procs of the WebSci10: Extending the frontiers of society on-line*, 2010.
http://richard.cyganiak.de/2008/papers/lod-usage-websci2010.pdf

Pastor-Sánchez, Juan-Antonio; **Martínez-Méndez, Francisco Javier**; **Rodríguez-Muñoz, José-Vicente**. "Advantages of thesaurus representation using the *simple knowledge organization system* (*SKOS*) compared with proposed alternatives". *Information research*, 2009, v. 14, n. 4, paper 422.
http://InformationR.net/ir/14-4/paper422.html